Some Bounds on Random Entropy Numbers with an Application to Support Vector Machines

Ingo Steinwart
Information Sciences Group, CCS-3
Mail Stop B256
Los Alamos National Laboratory
Los Alamos, NM 87545
Tel: (505) 665-7914
Fax: (505) 667-1126
ingo@lanl.gov

June 30, 2008

Abstract

In this paper we establish bounds on random entropy numbers of balls of reproducing kernel Hilbert spaces in terms of both eigenvalues of the associated integral operator and deterministic entropy numbers. We then use this bound to establish a new oracle inequality for support vector machines and show that this new oracle inequality is superior over existing oracle inequalities.

AMS subject classification
primary: 68Q32,
secondary: 15A18, 41A46, 47B06, 68T05

1 Introduction

Recent results [2, 16, 18] establishing learning rates for support vector machines (SVMs) use Talagrand's inequality together with local Rademacher averages, see [1], to bound the estimation error, i.e., the statistical error of these learning methods. This approach requires to bound the local Rademacher averages of relatively complicated function classes that depend on both the loss function and the reproducing kernel Hilbert space (RKHS) used in the SVM. For this task, two approaches exists: The first one, which goes back to [6] and is applied in [16, 18], uses Dudley's chaining together with $\|\cdot\|_\infty$-covering numbers of the RKHS, while the second one, applied in [2], uses [8] to bound the Rademacher averages by the the eigenvalues of the integral operator associated to the kernel of the RKHS. Both approaches have advantages and disadvantages. For example, compared to the $\|\cdot\|_\infty$-covering numbers, the eigenvalues are closer related to the learning problem at hand and provide, in general, a weaker notion of the complexity of the RKHS. In particular, the compactness of the input space is, in general, superfluous when using eigenvalues instead of $\|\cdot\|_\infty$-covering numbers. On the other hand, the analysis based on the eigenvalues is substantially more involved, and so far it is unclear whether apart from a relatively simple case considered in [2] it can be carried out for, e.g., more general loss functions. In addition, it remains so far unclear whether the analysis based on eigenvalues produces artifacts, such as the need of a quite restrictive noise assumption on the data-generating distribution and different learning rates for different exponents of the regularization term. As a result, both approaches cannot, so far, be compared on a fair ground, and it is thus unclear under which circumstances one or the other is superior.

In this paper, we address these issues by presenting a new technique for bounding the local Rademacher averages, which combines the advantages of both approaches and simultaneously lacks their disadvantages. At the heart of our approach lies the simple observation that one can use entropy numbers, which are the inverse concept of covering numbers, in Dudley's chaining argument. As a result (see Theorems 4.2 and 4.3), one can then bound the local Rademacher averages by the expectation of random entropy numbers. In the past, these in turn have been bound by $\|\cdot\|_\infty$-entropy (or covering) numbers which led to the first approach discussed above. To overcome the disadvantages of this approach, our main result stated in Theorem 2.1 shows that these random entropy numbers can be bounded by either the eigenvalues of the associated integral operator or related, deterministic entropy numbers. We then illustrate in Section 3 how this new bound can be used in the statistical analysis of SVMs. To this end, we establish a new oracle inequality for SVMs and derive learning rates from this inequality. By comparing these learning rates with the results found in [2], we then see that this new oracle inequality indeed combines the advantages of the two approaches discussed above while simultaneously lacking their disadvantages. In particular, it turns out that some of the requirements and findings of [2] are artifacts from their proof technique.

The rest of this paper is organized as follows. In Section 2 we recall some facts about RKHSs and eigenvalues. We then present our main result that bounds certain average entropy numbers associated to an RKHS by the eigenvalues of the corresponding integral operator of its kernel. Section 3 then shows how this bound can be used in the analysis of SVMs. Finally, Section 4 contains all proofs.

2 Bounds on Random Entropy Numbers

Let us begin by introducing some notations. To this end let H_1 and H_2 be two (real) Hilbert spaces and $S : H_1 \to H_2$ be a bounded linear operator. We denote the adjoint of S by S^*, i.e., S^* is the operator which is uniquely determined by the relation

$$\langle Sx, y\rangle_{H_2} = \langle x, S^*y\rangle_{H_1}, \qquad\qquad x \in H_1,\, y \in H_2.$$

Recall that an operator $T \in \mathcal{L}(H)$ is called self-adjoint if $T^* = T$, and it is called positive if $\langle Tx, x\rangle \geq 0$. Moreover, if the latter inequality is strict for all $x \neq 0$, we say that T is strictly positive. Given an $S \in \mathcal{L}(H_1, H_2)$, it is elementary to see that S^*S and SS^* are self-adjoint and positive.

A bounded operator $S : H_1 \to H_2$ is called compact if the closure of the image SB_{H_1}, where $B_{H_1} := \{x \in H_1 : \|x\|_{H_1} \leq 1\}$ denotes the closed unit ball of H_1, is a compact subset of H_2. One classical way to "quantify" the notion of compactness is to consider the (dyadic) entropy numbers which, for $i \geq 1$, are defined by

$$e_i(S) \ := \ \inf\biggl\{\varepsilon > 0 : \exists\, x_1, \dots, x_{2^{i-1}} \in H_1 \text{ such that } SB_{H_1} \subset \bigcup_{j=1}^{2^{i-1}} \bigl(x_j + \varepsilon B_{H_2}\bigr)\biggr\}.$$

Clearly, S is compact if and only if $\lim_{i\to\infty} e_i(S) = 0$, and the speed of this convergence can be considered as a measure how compact S is.

A $\lambda \in \mathbb{R}$ is an eigenvalue of $T \in \mathcal{L}(H)$ if there exists an $x \neq 0$ such that $Tx = \lambda x$. Every such x is called an eigenvector of T and λ. It is well-known that for compact, self-adjoint, and positive operators $T : H \to H$ there exist an at most countable orthonormal system $(e_i)_{i\in I}$ of H and a family $(\lambda_i(T))_{i\in I}$ such that $\lambda_1 \geq \lambda_2 \geq \dots > 0$ and

$$Tx = \sum_{i\in I} \lambda_i(T) \langle x, e_i\rangle e_i, \qquad\qquad x \in H. \tag{1}$$

Moreover, $\{\lambda_i(T) : i \in I\}$ is the set of non-zero eigenvalues of T. In the following, we assume that I is of the form $I = \{1, 2, \dots, |I|\}$ if the cardinality $|I|$ of I is finite. In this case, we define $\lambda_i(T) := 0$ for all $i > |I|$. Moreover, if $|I| = \infty$, we assume throughout this paper that $I = \mathbb{N}$. In the following, we call $(\lambda_i(T))_{i\geq 1}$ the extended sequence of eigenvalues of T.

Let us now recall some basic facts about reproducing kernel Hilbert spaces (RKHSs) which can be found in, e.g., Chapter 4 of [13]. To this end, let X be a non-empty set and H be a Hilbert space that consists of functions $f : X \to \mathbb{R}$. Then H is called an RKHS, if the Dirac functionals $\delta_x : H \to \mathbb{R}$, defined by $\delta_x(f) := f(x)$, are bounded linear operators for all $x \in X$. It is well-known that every RKHS has a unique representing kernel, i.e., a function $k : X \times X \to \mathbb{R}$ that satisfies $k(\,\cdot\,, x) \in H$ and

$$f(x) = \langle f, k(\,\cdot\,, x)\rangle_H\,, \qquad\qquad f \in H,\ x \in X.$$

Moreover, if k is measurable with respect to some σ-algebra $\mathcal{A}$ on X, then it is easy to show that H consists of measurable functions. Let us now assume that H is separable and that μ is a probability measure on $(X, \mathcal{A})$ such that

$$\|k\|_{L_2(\mu)} := \biggl(\int_X k(x,x)\, d\mu(x)\biggr)^{1/2} < \infty\,.$$

Then it is well-known that H consists of square integrable functions and the inclusion $\mathrm{id} : H \to L_2(\mu)$ is continuous with $\|\mathrm{id} : H \to L_2(\mu)\| \leq \|k\|_{L_2(\mu)}$. Moreover, the adjoint of this inclusion is the operator $S_{k,\mu} : L_2(\mu) \to H$ defined by

$$S_{k,\mu} g(x) := \int_X k(x,x') g(x') d\mu(x')\,, \qquad\qquad g \in L_2(\mu),\ x \in X. \tag{2}$$

In other words, we have $(\mathrm{id} : H \to L_2(\mu)) = S_{k,\mu}^*$. Furthermore, the integral operator $T_{k,\mu} := S_{k,\mu}^* \circ S_{k,\mu} : L_2(\mu) \to L_2(\mu)$ turns out to be self-adjoint, positive, and compact. In addition, its extended sequence of eigenvalues is summable, i.e.,

$$\sum_{i=1}^{\infty} \lambda_i(T_{k,\mu}) < \infty\,.$$

Let us now assume that we have a $D := (x_1, \dots, x_n) \in X^n$. Then the corresponding empirical measure is $\mathrm{D} := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, where δ_x denotes the Dirac measure at x, i.e., $\delta_x(A) = 1$ if $x \in A$ and $\delta_x(A) = 0$ otherwise. Clearly, D is a probability measure, and hence we can consider the embedding $\mathrm{id} : H \to L_2(\mathrm{D})$. The following theorem relates the entropy numbers of $\mathrm{id} : H \to L_2(\mathrm{D})$ to the entropy numbers of the embedding $\mathrm{id} : H \to L_2(\mu)$ if D is sampled from the product measure μ^n.

Theorem 2.1 *Let k be a measurable kernel on X with separable RKHS H and μ be a probability measure on X such that $\|k\|_{L_2(\mu)} < \infty$. Then for all $0 < p < \infty$ and all $0 < q \le 2$ there exists a constant $c_{p,q} \ge 1$ only depending on p and q such that for all $n \ge 1$, $m \ge 1$, and $M := \min\{m, n\}$ we have*

$$\sum_{i=1}^m i^{q/p-1}\mathbb{E}_{D\sim\mu^n} e_i^q(\mathrm{id} : H \to L_2(\mathrm{D})) \le c_{p,q} \sum_{i=1}^M i^{q/p-1}\biggl(\frac{1}{i}\sum_{j=i}^\infty \lambda_j(T_{k,\mu})\biggr)^{q/2}$$

and

$$\sum_{i=1}^m i^{q/p-1}\mathbb{E}_{D\sim\mu^n} e_i^q(\mathrm{id} : H \to L_2(\mathrm{D})) \le c_{p,q} \sum_{i=1}^M i^{q/p-1}\biggl(\frac{1}{i}\sum_{j=i}^\infty e_j^2\bigl(\mathrm{id} : H \to L_2(\mu)\bigr)\biggr)^{q/2}.$$

Let us illustrate the above theorem by the following two corollaries. The first corollary is based on Theorem 2.1 in the case of $q := 1$.

Corollary 2.2 *Let k be a measurable kernel on X with separable RKHS H and μ be a probability measure on X such that $\|k\|_{L_2(\mu)} < \infty$. Assume that there exist constants $0 < p < 2$ and $a \ge 1$ such that*

$$\lambda_i(T_{k,\mu}) \le a^2\, i^{-\frac{2}{p}}\,, \qquad\qquad i \ge 1, \tag{3}$$

or

$$e_i(\mathrm{id} : H \to L_2(\mu)) \le a\, i^{-\frac{1}{p}}\,, \qquad\qquad i \ge 1, \tag{4}$$

Then there exists a constant $c_p > 0$ only depending on p such that

$$\mathbb{E}_{D\sim\mu^n} e_i(\mathrm{id} : H \to L_2(\mathrm{D})) \le c_p\, a\, i^{-\frac{1}{p}}\,, \qquad\qquad i, n \ge 1.$$

Note that it is well-known that conditions (3) and (4) are actually *equivalent* modulo constants only depending on p. Indeed, by combining (14), (13), (16), and (18) we obtain $\lambda_i(T_{k,\mu}) \le 4e_i^2(\mathrm{id} : H \to L_2(\mu))$ for all $i \ge 1$. Conversely, Carl's inequality (17) together with (16), (13), and (14) can be used to show that (3) implies (4). Finally note that Theorem 3.4.2 in [3] shows a general equivalence between the behavior of the eigenvalues of $T_{k,\mu}$ and the entropy numbers of $\mathrm{id} : H \to L_2(\mu)$. For faster decaying sequences, however, this equivalence is more complicated.

In terms of Lorenz sequence spaces $\ell_{p,q}$, see Chapter 1.5 in [3], Corollary 2.2 states that $(e_i(\mathrm{id} : H \to L_2(\mu)))_{i\ge1} \in \ell_{p,\infty}$ for some $0 < p < 2$ implies

$$\bigl(\mathbb{E}_{D\sim\mu^n} e_i(\mathrm{id} : H \to L_2(\mathrm{D}))\bigr)_{i\ge1} \in \ell_{p,\infty}\,.$$

The next corollary provides an analogous implication for $\ell_{p,2}$.

Corollary 2.3 *Let k be a measurable kernel on X with separable RKHS H and μ be a probability measure on X such that $\|k\|_{L_2(\mu)} < \infty$. Then for all $0 < p < 2$ there exists a constant $c_p \geq 1$ only depending on p such that for all $n \geq 1$ we have*

$$\sum_{i=1}^{\infty} i^{2/p-1} \mathbb{E}_{D\sim\mu^n} e_i^2(\mathrm{id}: H \to L_2(\mathrm{D})) \leq c_p \sum_{i=1}^{\infty} i^{2/p-1} e_i^2(\mathrm{id}: H \to L_2(\mu))\,.$$

3 An Application to Support Vector Machines

Throughout this section we assume that X is a measurable space, $Y \subset [-1,1]$ is non-empty and compact, and P is a probability measure on $X \times Y$. Moreover, let H be a separable RKHS with bounded measurable kernel k satisfying $\|k\|_\infty \leq 1$. In addition, $L: Y \times \mathbb{R} \to [0,\infty)$ always denotes a continuous function which is convex in the second variable and satisfies $L(y,0) \leq 1$ for all $y \in Y$. Moreover, we assume that L is Lipschitz continuous in the sense of

$$\big|L(y,t_1) - L(y,t_2)\big| \leq |t_1 - t_2|\,, \qquad\qquad y \in Y,\, t_1, t_2 \in \mathbb{R}. \tag{5}$$

In particular, we are interested in the hinge loss, which for $Y := \{-1,1\}$ is defined by $L(y,t) := \max\{0, 1-yt\}$, $y \in Y$, $t \in \mathbb{R}$. The function L will serve as loss function and consequently let us recall the associated L-risk

$$\mathcal{R}_{L,P}(f) := \mathbb{E}_{(x,y)\sim P} L(y, f(x)),$$

where $f: X \to \mathbb{R}$ is a measurable function. Note that our assumptions immediately give $\mathcal{R}_{L,P}(0) \leq 1$. Furthermore, the minimal L-risk is denoted by $\mathcal{R}^*_{L,P}$, i.e.

$$\mathcal{R}^*_{L,P} := \inf\{\mathcal{R}_{L,P}(f) \,|\, f: X \to \mathbb{R} \text{ measurable}\},$$

and a function attaining this infimum is denoted by $f^*_{L,P}$. In the following we always assume that there exists at least one such $f^*_{L,P}$.

Recall that support vector machines, see [4, 10, 13], are based on the optimization

$$f_{P,\lambda} := \arg\min_{f\in H} \Big(\lambda \|f\|_H^2 + \mathcal{R}_{L,P}(f)\Big)\,,$$

where $\lambda > 0$ is a user-defined regularization parameter and the function $f_{P,\lambda}$ is known to be uniquely determined. Note that if we identify a training set $D = ((x_1,y_1),\dots,(x_n,y_n)) \in (X \times Y)^n$ with its empirical measure, then $f_{D,\lambda}$ denotes the empirical estimators of the above learning scheme.

One way to describe the approximation error of support vector machines is the approximation error function

$$A(\lambda) := \lambda \|f_{P,\lambda}\|_H^2 + \mathcal{R}_{L,P}(f_{P,\lambda}) - \mathcal{R}^*_{L,P}\,, \qquad\qquad \lambda > 0,$$

which is discussed in some detail in [15] and Chapter 5.4 of [13].

Theorem 3.1 *Let L, H, and P satisfy the above assumptions. Moreover, assume that there are constants $a \geq 1$ and $0 < p < 2$ such that*

$$\lambda_i(T_{k,P_X}) \leq a^{\frac{2}{p}}\, i^{-\frac{2}{p}}\,, \qquad\qquad i \geq 1, \tag{6}$$

In addition, suppose that for all $0 < \lambda \leq 1$ *and all* $f \in \lambda^{-\frac{1}{2}} B_H$ *we have*

$$\mathbb{E}_P \big(L \circ f - L \circ f^*_{L,P}\big)^2 \leq c \left(\|f\|_\infty + 1\right)^v \left(\mathbb{E}_P L \circ f - L \circ f^*_{L,P}\right)^\vartheta \tag{7}$$

for some constants $c \geq 1$*,* $\vartheta \in (0, 1]$*, and* $v \in [0, 2]$*. Then there exists a constant* $K \geq 1$ *such that for all* $0 < \lambda \leq 1$*,* $\varepsilon \in (0, 1]$*,* $x \geq 1$ *satisfying*

$$\begin{aligned}\varepsilon \quad \geq \quad \max\Bigg\{ & A(\lambda) + \lambda, \left(\frac{Ka}{\lambda^{\frac{2\alpha p + v(2-p)}{4}} n}\right)^{\frac{4}{8 - 2\alpha p - (v + 2\vartheta)(2-p)}}, \left(\frac{Ka}{\lambda^{\frac{\alpha(2+p)}{4}} n}\right)^{\frac{4}{(2+p)(2-\alpha)}}, \\ & \left(\frac{Kx}{\lambda^{\frac{v}{2}} n}\right)^{\frac{2}{4 - v - 2\vartheta}}, \left(\frac{Kx}{\lambda^{\frac{\alpha}{2}} n}\right)^{\frac{2}{2-\alpha}} \Bigg\},\end{aligned}$$

we have

$$P^n \Big(D \in (X \times Y)^n : \mathcal{R}_{L,P}(f_{D,\lambda}) - \mathcal{R}^*_{L,P} < A(\lambda) + \varepsilon \Big) \; \geq \; 1 - e^{-x} \,.$$

In order to illustrate the above theorem let us now assume that L is the hinge loss introduced above. Moreover, assume that P is a distribution with Tsybakov noise exponent $q \in [0, \infty]$, i.e., there exists a $C > 0$ such that for $\eta(x) := P(y = 1|x)$, $x \in X$, and all sufficiently small $t > 0$ we have

$$P_X\big(\{x \in X : |2\eta(x) - 1| \leq t\}\big) \; \leq \; C \cdot t^q \,.$$

When $q > 0$, it follows from [16, Lemma 6.6] that the assumption (7) is satisfied with $\alpha = 1$, $v = \frac{q+2}{q+1}$, $\vartheta = \frac{q}{q+1}$ and $C = \|(2\eta - 1)^{-1}\|_{q,\infty} + 2$. Moreover it is simple to show the same is true when $q = 0$ but with $C = 5$. Consequently, the condition on ε becomes

$$\varepsilon \quad \geq \quad \max\bigg\{ A(\lambda) + \lambda, \frac{K}{\lambda}\Big(\frac{a}{n}\Big)^{\frac{4(q+1)}{2q + pq + 4}}, \frac{K}{\lambda}\Big(\frac{a}{n}\Big)^{\frac{4}{2+p}}, \frac{K}{\lambda}\Big(\frac{x}{n}\Big)^{\frac{2(q+1)}{q+2}}, \frac{K}{\lambda}\Big(\frac{x}{n}\Big)^2 \bigg\}.$$

Some easy estimates then show that this reduces to

$$\varepsilon \geq A(\lambda) + \lambda + K x^2 \lambda^{-1} \Big(\frac{a}{n}\Big)^{\frac{4(q+1)}{2q + pq + 4}} \,, \tag{8}$$

where $K \geq 1$ is a suitable constant and a and n are assumed to satisfy $n \geq a \geq 1$. Let us now assume that there exists constants $c > 0$ and $\beta \in (0, 1]$ such that $A(\lambda) \leq c\lambda^\beta$ for all $\lambda > 0$. For $\lambda_n := n^{-\frac{4(q+1)}{(2q+pq+4)(1+\beta)}}$ estimate (8) then immediately yields the learning rate

$$n^{-\frac{4\beta(q+1)}{(2q+pq+4)(1+\beta)}} \,,$$

which has already been established in [15, Theorem 1] for continuous kernels over compacta under the more restrictive entropy number assumption

$$e_i(\mathrm{id} : H \to C(X)) \leq a^{\frac{1}{p}} \, i^{-\frac{1}{p}} \,, \qquad\qquad i \geq 1. \tag{9}$$

Let us now compare the learning rate above to the findings of [2]. To this end, it suffices to consider the case $q = \infty$, i.e., the conditional probability $\eta(x)$ is bounded away from the critical level 1/2. In this case our learning rate reduces to

$$n^{-\frac{4\beta}{(2+p)(1+\beta)}} \,. \tag{10}$$

On the other hand, for

$$\gamma(n) := \frac{1}{\sqrt{n}} \inf_{j\geq 1} \left(\frac{j}{\sqrt{n}} + \sqrt{\sum_{i\geq j} \lambda_i(T_{k,P_X})} \right), \qquad n \geq 1,$$

assumption (6) yields a constant $c_{a,p}$ such that $\gamma(n) \leq c_{a,p} n^{-\frac{2}{2+p}}$ for all $n \geq 1$. The oracle inequality for SVMs derived in [2, Theorem 3.1] thus yields the rate

$$n^{-\frac{2\beta}{2+p}}, \tag{11}$$

which is obviously worse by a factor of $\frac{2}{1+\beta}$ in the exponent. Moreover, note that this oracle inequality required additionally, that η is also bounded away from the most benign levels 0 and 1, whereas our result does not require this somewhat nonnatural assumption. Moreover, [2, Theorem 3.1] also provides an oracle inequality for SVMs that use the regularization term $\lambda\|f\|_H$ which is lighter than the standard $\lambda\|f\|_H^2$ term. In order to derive a learning rate from this oracle inequality we write

$$A^*(\lambda) := \inf_{f\in H} \Bigl(\lambda\|f\|_H + \mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P} \Bigr), \qquad \lambda > 0,$$

for the approximation error function that corresponds to this lighter regularization. Using the techniques from [15, Appendix] it is easy to show that our assumption $A(\lambda) \leq c\lambda^\beta$ implies $A^*(\lambda) \leq \tilde{c}\lambda^{\frac{2\beta}{1+\beta}}$ for a constant $\tilde{c}$ and all $\lambda > 0$. Consequently, the oracle inequality of [2, Theorem 3.1] yields the learning rate (10). Note that the discrepancy of the learning rates (10) and (11) make the authors in [2] conjecture that the lighter regularization may be superior. Of course, our analysis neither proves or disproves this conjecture since so far no lower bounds are known. However, our analysis shows at least that the techniques of [2] are sub-optimal for analyzing SVMs with standard regularization term.

4 Proofs

In this section we provide the proofs for the results of Section 2. To this end we recall some results on entropy numbers and their relation to eigenvalues in Subsection 4.1. The proofs of the main results are then presented in Subsection 4.2.

4.1 Eigenvalues, Singular Numbers, and Entropy Numbers

Given a compact operator $S : H_1 \to H_2$, the operator $S^*S : H_1 \to H_1$ is compact, positive, and self-adjoint, and hence it enjoys a representation of the form (1) with non-negative eigenvalues. We write

$$s_i(S) := \sqrt{\lambda_i(S^*S)}, \qquad i \geq 1 \tag{12}$$

for the singular numbers of S, where $(\lambda_i(S^*S))_{i\geq 1}$ is the extended sequence of eigenvalues of S^*S. Recall that S^*S and SS^* have exactly the same non-zero eigenvalues with the same geometric multiplicities and hence we find $s_i(S^*) = s_i(S)$ for all $i \geq 1$. Moreover, we have

$$s_i^2(S) = \lambda_i(S^*S) = s_i(S^*S)\,, \qquad\qquad i \geq 1, \tag{13}$$

where in the second equality we used the fact that for compact, positive, and self-adjoint $T : H \to H$ we have

$$s_i(T) = \sqrt{\lambda_i(T^*T)} = \sqrt{\lambda_i(T^2)} = \lambda_i(T)\,, \qquad\qquad i \geq 1. \tag{14}$$

Let us now consider another interesting property of the singular numbers. To this end, let $S \in \mathcal{L}(E, F)$ be a bounded operator acting between arbitrary Banach spaces E and F. For $i \geq 1$, its i-th approximation number is defined by

$$a_i(S) := \inf\bigl\{\|S - A\| : A \in \mathcal{L}(E, F) \text{ with rank } A < i\bigr\}\,. \tag{15}$$

Obviously, $(a_i(S))_{i\geq1}$ is decreasing, and if rank $S < \infty$, we also have $a_i(S) = 0$ for all $i >$ rank S. Moreover, by diagonalization (see, e.g., Section 2.11 of [9]), one can show that

$$s_i(S) = a_i(S) \tag{16}$$

for all compact $S \in \mathcal{L}(H_1, H_2)$ acting between Hilbert spaces and all $i \geq 1$.

Entropy numbers are closely related to the approximation numbers introduced in Namely, Carl's inequality, see Theorem 3.1.2 in [3], states that for all $0 < p \leq \infty$ and $0 < q < \infty$ there exists a constant $c_{p,q} > 0$ such that

$$\sum_{i=1}^{m} i^{q/p-1} e_i^q(S) \leq c_{p,q} \sum_{i=1}^{m} i^{q/p-1} a_i^q(S) \tag{17}$$

for all bounded operators $S : E \to F$ acting between Banach spaces and all $m \geq 1$. Moreover, for Hilbert spaces H and compact operators $T : H \to H$, we have the following strong inverse of the above inequality:

$$a_i(T) \leq 2e_i(T)\,, \qquad\qquad i \geq 1\,. \tag{18}$$

For a proof we refer to p. 120 in [3].

4.2 Proofs of the Random Entropy Number Bounds

Besides the material from Subsection 4.1 we also need the following result on random eigenvalues for the proof of Theorem 2.1. This result was first shown by [11, 12] in the special case of continuous kernels over compact metric spaces, and [19] generalized this result to bounded measurable kernels with separable RKHSs. However, a close inspection of the proof in [19] shows that the boundedness of the kernel k can be replaced by the weaker assumption $\|k\|_{L_2(\mu)} < \infty$.

Theorem 4.1 *Let k be a measurable kernel on X with separable RKHS H and μ be a probability measure on X such that $\|k\|_{L_2(\mu)} < \infty$. Then for all $m \geq 1$ we have*

$$\mathbb{E}_{D\sim\mu^n} \sum_{i=m}^{\infty} \lambda_i(T_{k,\mathrm{D}}) \leq \sum_{i=m}^{\infty} \lambda_i(T_{k,\mu})\,. \tag{19}$$

Proof of Theorem 2.1: Carl's inequality (17) shows that there exists a constant $c_{p,q} > 0$ such that for $m, n \geq 1$ and all $D \in X^n$ we have

$$\sum_{i=1}^{m} i^{q/p-1} e_i^q(S^*_{k,\mathrm{D}}) \leq c_{p,q} \sum_{i=1}^{m} i^{q/p-1} a_i^q(S^*_{k,\mathrm{D}}) = c_{p,q} \sum_{i=1}^{\min\{m,n\}} i^{q/p-1} a_i^q(S^*_{k,\mathrm{D}})\,,$$

where in the last step we used that $n \geq \operatorname{rank} S^*_{k,D}$ implies $a_i(S^*_{k,\mathrm{D}}) = 0$ for all $i > n$. Moreover, for $M := \min\{m, n\}$ and $\tilde{M} := \lfloor (M+1)/2 \rfloor$, we have

$$\sum_{i=1}^{M} i^{q/p-1} a_i^q(S^*_{k,\mathrm{D}}) \leq \sum_{i=1}^{\tilde{M}} (2i-1)^{q/p-1} a_{2i-1}^q(S^*_{k,\mathrm{D}}) + \sum_{i=1}^{\tilde{M}} (2i)^{q/p-1} a_{2i}^q(S^*_{k,\mathrm{D}})\,.$$

If $p \leq q$, the monotonicity of the approximation numbers thus yields

$$\sum_{i=1}^{M} i^{q/p-1} a_i^q(S^*_{k,\mathrm{D}}) \leq 2^{q/p} \sum_{i=1}^{M} i^{q/p-1} a_{2i-1}^q(S^*_{k,\mathrm{D}})\,,$$

and if $p > q$ we analogously find

$$\sum_{i=1}^{M} i^{q/p-1} a_i^q(S^*_{k,\mathrm{D}}) \leq 2^{2+q/p} \sum_{i=1}^{M} i^{q/p-1} a_{2i-1}^q(S^*_{k,\mathrm{D}})\,,$$

Using (16) and (12) we further see that $a_i^2(S^*_{k,\mathrm{D}}) = s_i^2(S^*_{k,\mathrm{D}}) = s_i(S^*_{k,\mathrm{D}} S_{k,\mathrm{D}}) = \lambda_i(T_{k,\mathrm{D}})$ for all $i \geq 1$ and $D \in X^n$. Since $q \leq 2$ we thus obtain

$$\begin{aligned} \sum_{i=1}^{m} i^{q/p-1} \mathbb{E}_{D\sim\mu^n} e_i^q(S^*_{k,\mathrm{D}}) &\leq \tilde{c}_{p,q} \sum_{i=1}^{M} i^{q/p-1} \mathbb{E}_{D\sim\mu^n} a_{2i-1}^q(S^*_{k,\mathrm{D}}) \\ &\leq \tilde{c}_{p,q} \sum_{i=1}^{M} i^{q/p-1} \bigl(\mathbb{E}_{D\sim\mu^n} \lambda_{2i-1}(T_{k,\mathrm{D}})\bigr)^{q/2}\,, \end{aligned}$$

where $\tilde{c}_{p,q} := 2^{2+q/p} c_{p,q}$. Now for each $D \in X^n$ the sequence $(\lambda_i(T_{k,\mathrm{D}}))_{i\geq 1}$ is monotonically decreasing and hence so is $(\mathbb{E}_{D\sim\mu^n} \lambda_i(T_{k,\mathrm{D}}))_{i\geq 1}$. By Theorem 4.1, we hence find

$$i\,\mathbb{E}_{D\sim\mu^n} \lambda_{2i-1}(T_{k,\mathrm{D}}) \leq \sum_{j=i}^{2i-1} \mathbb{E}_{D\sim\mu^n} \lambda_j(T_{k,\mathrm{D}}) \leq \sum_{j=i}^{\infty} \lambda_j(T_{k,\mu})$$

for all $i \geq 1$, and consequently we obtain

$$\sum_{i=1}^{M} i^{q/p-1} \bigl(\mathbb{E}_{D\sim\mu^n} \lambda_{2i-1}(T_{k,\mathrm{D}})\bigr)^{q/2} \leq \sum_{i=1}^{M} i^{q/p-1} \biggl(\frac{1}{i} \sum_{j=i}^{\infty} \lambda_j(T_{k,\mu})\biggr)^{q/2}\,.$$

Combining the estimates above, we then obtain the first assertion. Moreover, by (12), (13), and (16), we have

$$\lambda_j(T_{k,\mu}) = s_i(S^*_{k,\mu} \circ S_{k,\mu}) = s_i^2(S^*_{k,\mu}) = a_j^2(S^*_{k,\mu}) \leq 4e_j^2(S^*_{k,\mu})\,,$$

where in the last step we used (18). Combining the estimates above, we hence obtain

$$\sum_{i=1}^{m} i^{q/p-1} \mathbb{E}_{D\sim\mu^n} e_i^q(S^*_{k,\mathrm{D}}) \leq 2^q \tilde{c}_{p,q} \sum_{i=1}^{M} i^{q/p-1} \biggl(\frac{1}{i} \sum_{j=i}^{\infty} e_j^2(S^*_{k,\mu}) \biggr)^{q/2} ,$$

i.e., we have also shown the second assertion. ■

Proof of Corollary 2.2: Since $0 < p < 2$, it is easy to see that there exists a constant $\tilde{c}_p$ such that

$$\frac{1}{i} \sum_{j=i}^{\infty} \lambda_j(T_{k,\mu}) \leq a^2 \cdot \frac{1}{i} \sum_{j=i}^{\infty} j^{-\frac{2}{p}} \leq \tilde{c}_p^2 \, a^2 \, i^{-\frac{2}{p}}$$

for all $i \geq 1$. Using $\frac{1}{p} - 1 > -1$, we hence find another constant $c_p' > 0$ such that for $m \geq 1$ we have

$$\sum_{i=1}^{m} i^{\frac{2}{p}-1} \biggl(\frac{1}{i} \sum_{j=i}^{\infty} \lambda_j(T_{k,\mu}) \biggr)^{1/2} \leq \tilde{c}_p \, a \sum_{i=1}^{m} i^{\frac{1}{p}-1} \leq c_p' \, a \, m^{\frac{1}{p}} . \qquad (20)$$

Furthermore, for $\tilde{m} := \lfloor (m+1)/2 \rfloor$, the monotonicity of the entropy numbers yields

$$\tilde{m}^{\frac{2}{p}} \, \mathbb{E}_{D\sim\mu^n} e_m(S^*_{k,\mathrm{D}}) \leq \sum_{i=\tilde{m}}^{m} i^{\frac{2}{p}-1} \, \mathbb{E}_{D\sim\mu^n} e_i(S^*_{k,\mathrm{D}}) \leq \sum_{i=1}^{m} i^{\frac{2}{p}-1} \, \mathbb{E}_{D\sim\mu^n} e_i(S^*_{k,\mathrm{D}}) ,$$

and since $m/2 \leq \lfloor (m+1)/2 \rfloor = \tilde{m}$, we hence obtain

$$\mathbb{E}_{D\sim\mu^n} e_m(S^*_{k,\mathrm{D}}) \leq 4^{1/p} m^{-\frac{2}{p}} \sum_{i=1}^{m} i^{\frac{2}{p}-1} \, \mathbb{E}_{D\sim\mu^n} e_i(S^*_{k,\mathrm{D}}) .$$

Combining this estimate with (20) and Theorem 2.1 for $\tilde{p} := p/2$ and $q := 1$ then yields first assertion if (3) is satisfied. The second case can be shown completely analogously. ■

Proof of Corollary 2.3: For $q = 2$ the right-hand side of the second estimate of Theorem 2.1 becomes

$$\sum_{i=1}^{M} i^{q/p-1} \biggl(\frac{1}{i} \sum_{j=i}^{\infty} e_j^2(S^*_{k,\mu}) \biggr)^{q/2} = \sum_{i=1}^{M} i^{2/p-2} \sum_{j=i}^{\infty} e_j^2(S^*_{k,\mu}) = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} b_{i,j} ,$$

where $b_{i,j} := 0$ if $i > \min\{j, M\}$ and $b_{i,j} := i^{2/p-2} e_j^2(S^*_{k,\mu})$ otherwise. Moreover, rearranging the sums and using $p < 2$ yields a constant c_p only depending on p such that

$$\begin{aligned}
\sum_{j=1}^{\infty} \sum_{i=1}^{\infty} b_{i,j} &= \sum_{j=1}^{M} \sum_{i=1}^{j} i^{2/p-2} e_j^2(S^*_{k,\mu}) + \sum_{j=M+1}^{\infty} \sum_{i=1}^{M} i^{2/p-2} e_j^2(S^*_{k,\mu}) \\
&\leq c_p \sum_{j=1}^{M} j^{2/p-1} e_j^2(S^*_{k,\mu}) + c_p \sum_{j=M+1}^{\infty} M^{2/p-1} e_j^2(S^*_{k,\mu}) \\
&\leq c_p \sum_{j=1}^{\infty} j^{2/p-1} e_j^2(S^*_{k,\mu})
\end{aligned}$$

Applying Theorem 2.1 then yields the assertion. ■

4.3 Bounding Rademacher Averages by Random Entropy Numbers

Given a probability space $(\Theta, \mathcal{C}, \nu)$, let us recall that a finite sequence $\varepsilon_1, \dots, \varepsilon_n$ of i.i.d. random variables $\varepsilon_i : \Theta \to \{-1, 1\}$, $i = 1, \dots, n$, is called a Rademacher sequence if $\nu(\varepsilon_i = 1) = \nu(\varepsilon_i = -1) = 1/2$ for all $i = 1, \dots, n$. Now let Z be a non-empty set equipped with some σ-algebra and $\mathcal{L}_0(Z)$ be the corresponding set of all measurable functions $g : Z \to \mathbb{R}$. Given a non-empty $\mathcal{G} \subset \mathcal{L}_0(Z)$, a Rademacher sequence $\varepsilon_1, \dots, \varepsilon_n$, and a finite sequence $D := (z_1, \dots, z_n) \in Z^n$, the n-th empirical Rademacher average of $\mathcal{G}$ is defined by

$$\mathrm{Rad}_D(\mathcal{G}, n) := \mathbb{E}_\nu \sup_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i g(z_i) \right| .$$

With the help of Dudley's chaining argument, see [5] and Chapter 2.2 in [17], one can bound empirical Rademacher averages $\mathrm{Rad}_D(\mathcal{G}, n)$ by the covering numbers of $\mathcal{G}$ with respect to $L_2(\mathrm{D})$. We refer to [7] for an overview of this approach. However, for our purposes it is more convenient to use entropy numbers instead of covering numbers. Fortunately, Dudley's chaining argument works with entropy numbers as well as with covering numbers. In order to present the corresponding result, we define the (dyadic) entropy numbers of a subset $A \subset H$ of a Hilbert space H by

$$e_i(A, H) \ := \ \inf \left\{ \varepsilon > 0 : \exists x_1, \dots, x_{2^{i-1}} \in H \text{ such that } A \subset \bigcup_{j=1}^{2^{i-1}} \bigl(x_j + \varepsilon B_H \bigr) \right\} .$$

Note that for a bounded operator $S : H_1 \to H_2$ this definition yields $e_i(S) = e_i(SB_{H_1}, H_2)$ for all $i \geq 1$. Now the following result whose proof can be found in Chapter 7.3 of [13] bounds empirical Rademacher averages by entropy numbers.

Theorem 4.2 *For every non-empty set $\mathcal{G} \subset \mathcal{L}_0(Z)$ and every finite sequence $D := (z_1, \dots, z_n) \in Z^n$, we have*

$$\mathrm{Rad}_D(\mathcal{G}, n) \leq \sqrt{\frac{\ln 16}{n}} \left(\sum_{i=1}^\infty 2^{i/2}\, e_{2^i}\bigl(\mathcal{G} \cup \{0\}, L_2(\mathrm{D})\bigr) + \sup_{g \in \mathcal{G}} \|g\|_{L_2(\mathrm{D})} \right) .$$

The analysis of learning algorithms usually require to consider the expectation of empirical Rademacher averages. Using Theorem 4.2 and an imposed bound on the average entropy numbers the following theorem provides a bound on expected Rademacher averages. Its proof follows the ideas of [6] and can be found in Chapter 7.3 of [13].

Theorem 4.3 *Let $\mathcal{G} \subset \mathcal{L}_0(Z)$ be a non-empty set and P be a distribution on Z. Suppose that there exist constants $B \geq 0$ and $\sigma \geq 0$ such that $\|h\|_\infty \leq B$ and $\mathbb{E}_P h^2 \leq \sigma$ for all $h \in \mathcal{G}$. Furthermore, assume that for a fixed $n \geq 1$ there exist constants $p \in (0, 2)$ and $a \geq B^p$ such that*

$$\mathbb{E}_{D \sim P^n}\, e_i(\mathcal{G}, L_2(\mathrm{D})) \ \leq \ a^{\frac{1}{p}}\, i^{-\frac{1}{p}} , \qquad i \geq 1. \tag{21}$$

Then there exist constants $C_1(p) > 0$ and $C_2(p) > 0$ depending only on p such that

$$\mathbb{E}_{D \sim P^n} \mathrm{Rad}_D(\mathcal{G}, n) \leq \max\left\{ C_1(p)\, a^{\frac{1}{2}} \sigma^{\frac{2-p}{4}} n^{-\frac{1}{2}},\, C_2(p)\, a^{\frac{2}{2+p}} B^{\frac{2-p}{2+p}} n^{-\frac{2}{2+p}} \right\} .$$

4.4 Bounding Rademacher Averages by Random Entropy Numbers

Let us begin by introducing some additional notations. To this end, let $L : Y \times \mathbb{R} \to [0, \infty)$ be continuous function which is convex in the second variable and satisfies $L(y, 0) \le 1$ for all $y \in Y$. Moreover, we assume that L is Lipschitz continuous in the sense of (5). In addition, let H be a separable RKHS over X with bounded measurable kernel k satisfying $\|k\|_\infty \le 1$, and P be a probability measure on $X \times Y$. Given a fixed $\lambda > 0$ we define

$$g_f := \lambda\|f\|_H^2 + L\circ f - \lambda\|f_{P,\lambda}\|_H^2 - L\circ f_{P,\lambda}, \qquad\qquad f \in H,$$

where $L \circ f$ denotes the function $(x, y) \mapsto L(y, f(x))$. Moreover, we need the set $\mathcal{G}(\lambda) := \{g_f : f \in \lambda^{-1/2} B_H\}$ and for fixed $\varepsilon > 0$ we further write

$$\mathcal{G}_\varepsilon := \{g \in \mathcal{G}(\lambda) : \mathbb{E}_P g \le \varepsilon\}$$

and

$$\Lambda := \left(\frac{A(\lambda)+\varepsilon}{\lambda}\right)^{1/2}.$$

For the proof of the following lemma we finally need to recall the elementary Lemma 4.1 in [14] which showed $\|f\|_H \le \Lambda$ for all $f \in H$ with $g_f \in \mathcal{G}_\varepsilon$.

Lemma 4.4 *Assume that the conditions on L and H mentioned above are true. Furthermore, let $n \in \mathbb{N}$, and assume that there are constants $a \ge 1$ and $p \in (0, 2)$ such that*

$$\mathbb{E}_{D_X\sim P_X^n}\, e_i(\mathrm{id} : H \to L_2(\mathrm{D}_X)) \;\le\; a^{\frac{1}{p}}\, i^{-\frac{1}{p}}, \qquad\qquad i \ge 1. \tag{22}$$

Then there is a constant $c_p > 0$ depending only on p such that for all distributions P on $X \times Y$, all $\lambda \in (0, 1]$, $\varepsilon \in (0, 1]$, and all $\tau_\varepsilon \ge \sup_{g\in\mathcal{G}_\varepsilon} \mathbb{E}_P g^2$ we have

$$\mathbb{E}_{D\sim P^n}\mathrm{Rad}_D(\mathcal{G}_\varepsilon, n) \le c_p\, \max\left\{(\Lambda^2+1)^{\frac{p}{4}}\tau_\varepsilon^{\frac{2-p}{4}}\left(\frac{a}{n}\right)^{\frac{1}{2}}, (\Lambda^2+1)^{\frac{1}{2}}\left(\frac{a}{n}\right)^{\frac{2}{2+p}}\right\}.$$

Proof: Let us write $\tilde{\mathcal{G}}_\varepsilon := \{\lambda\|f\|_H^2 + L\circ f : f \in \Lambda B_H\}$ and $\mathcal{H} := \{L\circ f : f \in \Lambda B_H\}$. Now observe that $\lambda\|f\|_H^2 \le 2$ for all $f \in \Lambda B_H$, and hence the additivity of the entropy numbers, see ..., together with the Lipschitz continuity of L yields

$$\begin{aligned} e_{2i-1}\big(\mathcal{G}_\varepsilon, L_2(\mathrm{D})\big) = e_{2i-1}\big(\tilde{\mathcal{G}}_\varepsilon, L_2(\mathrm{D})\big) \;&\le\; e_i\big([0,2], |\cdot|\big) + e_i\big(\mathcal{H}, L_2(\mathrm{D}_X)\big) \\ &\le\; 2^{1-i} + e_i\big(\Lambda B_H, L_2(\mathrm{D}_X)\big) \end{aligned}$$

for all $i \ge 1$ and all $D := ((x_1, y_1), \ldots, (x_n, y_n)) \in (X \times Y)^n$, where D_X is the empirical measure associated to $D_X := (x_1, \ldots, x_n)$. Averraging over D and using (22) we thus obtain

$$\mathbb{E}_{D\sim P^n} e_{2i-1}\big(\mathcal{G}_\varepsilon, L_2(\mathrm{D})\big) \le 2^{1-i} + \Lambda a^{\frac{1}{p}}\, i^{-\frac{1}{p}} \le \tilde{c}_p(\Lambda^2+1)^{\frac{1}{2}} a^{\frac{1}{p}}\, i^{-\frac{1}{p}}$$

for a suitable constant $\tilde{c}_p$ only depending on p. From this it is straightforward to conclude that

$$\mathbb{E}_{D\sim P^n} e_i\big(\mathcal{G}_\varepsilon, L_2(\mathrm{D})\big) \le c_p(\Lambda^2+1)^{\frac{1}{2}} a^{\frac{1}{p}}\, i^{-\frac{1}{p}}$$

for all $i \geq 1$, where c_p is another constant only depending on p. Now observe that, for $f \in H$ with $g_f \in \mathcal{G}_\varepsilon$, we have $\|L \circ f\|_\infty \leq 1 + \|f\|_\infty \leq 1 + \|f\|_H \leq 1 + \Lambda$ and $\lambda \|f\|_H^2 \leq 2$. From this it is easy to conclude that $\|g_f\|_\infty \leq \Lambda + 3 =: B$ for all $g_f \in \mathcal{G}_\varepsilon$. Assuming without loss of generality that $c_p \geq \sqrt{10}$ we hence find for $\tilde{a} := c_p^p (\Lambda^2 + 1)^{\frac{p}{2}} a$ that $\tilde{a} \geq B^p$. Applying Theorem 4.3 then yields the assertion. ■

Proof of Theorem 3.1: We first note that, besides the assumed entropy number bound, Theorem 3.1 is a slightly simplified version of Theorem 2.1 of [14]. An inspection of the proof of Theorem 2.1 of [14] shows that the entropy number assumption imposed in that theorem is only used in Lemma 4.3 of [14] to bound the modulus of continuity $\omega_{P,n}(\mathcal{G}(\lambda), \varepsilon)$. Moreover, symmetrization yields $\omega_{P,n}(\mathcal{G}(\lambda), \varepsilon) \leq 2\mathbb{E}_{D \sim P^n} \mathrm{Rad}_D(\mathcal{G}_\varepsilon, n)$, and using Corollary 2.2, we can hence replace Lemma 4.3 of [14] by Lemma 4.4 above. The rest of the proof of Theorem 2.1 of [14] is not affected. ■

References

[1] P.L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *Ann. Statist.*, 33:1497–1537, 2005.

[2] G. Blanchard, O. Bousquet, and P. Massart. Statistical performance of support vector machines. *Ann. Statist.*, 36:489–531, 2008.

[3] B. Carl and I. Stephani. *Entropy, Compactness and the Approximation of Operators.* Cambridge University Press, 1990.

[4] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines.* Cambridge University Press, 2000.

[5] R. M. Dudley. The sizes of compact subsets of Hilbert spaces and continuity of Gaussian processes. *J. Funct. Anal.*, 1:290–330, 1967.

[6] S. Mendelson. Improving the sample complexity using global data. *IEEE Trans. Inform. Theory*, 48:1977–1991, 2002.

[7] S. Mendelson. A few notes on statistical learning theory. In S. Mendelson and A. Smola, editors, *Advanced Lectures on Machine Learning: Machine Learning Summer School 2002, Canberra, Australia*, pages 1–40. Springer, Berlin, 2003.

[8] S. Mendelson. On the performance of kernel classes. *J. Mach. Learn. Res.*, 4:759–771, 2003.

[9] A. Pietsch. *Eigenvalues and s-Numbers.* Geest & Portig K.-G., Leipzig, 1987.

[10] B. Schölkopf and A.J. Smola. *Learning with Kernels.* MIT Press, 2002.

[11] J. Shawe-Taylor, C. K I. Williams, N. Cristianini, and J. Kandola. On the eigenspectrum of the Gram matrix and its relationship to the operator eigenspectrum. In N. Cesa-Bianchi, M. Numao, and R. Reischuk, editors, *Algorithmic Learning Theory, 13th International Conference*, pages 23–40. Springer, New York, 2002.

[12] J. Shawe-Taylor, C. K. I. Williams, N. Cristianini, and J. Kandola. On the eigenspectrum of the Gram matrix and the generalization error of kernel-PCA. *IEEE Trans. Inform. Theory*, 51:2510–2522, 2005.

[13] I. Steinwart and A. Christmann. *Support Vector Machines.* Springer, New York, 2008.

[14] I. Steinwart, D. Hush, and C. Scovel. A new concentration result for regularized risk minimizers. In E. Giné, V. Koltchinskii, W. Li, and J. Zinn, editors, *High Dimensional Probability IV*, volume 51 of *Lecture Notes–Monograph Series*, pages 260–275, Beachwood, 2006. Institute of Mathematical Statistics.

[15] I. Steinwart and C. Scovel. Fast rates for support vector machines. In *Proceedings of the 18th Annual Conference on Learning Theory, COLT 2005*, pages 279–294. Springer, 2005.

[16] I. Steinwart and C. Scovel. Fast rates for support vector machines using Gaussian kernels. *Ann. Statist.*, 35:575–607, 2007.

[17] A.W. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes.* Springer, New York, 1996.

[18] Q. Wu, Y. Ying, and D.-X. Zhou. Multi-kernel regularized classifiers. *J. Complexity*, 23:108–134, 2007.

[19] L. Zwald, O. Bousquet, and G. Blanchard. Statistical properties of kernel principal component analysis. In J. Shawe-Taylor and Y. Singer, editors, *Proceedings of the 17th Annual Conference on Learning Theory*, pages 594–608. Springer, New York, 2004.